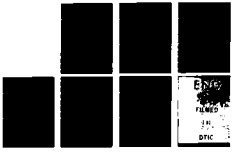
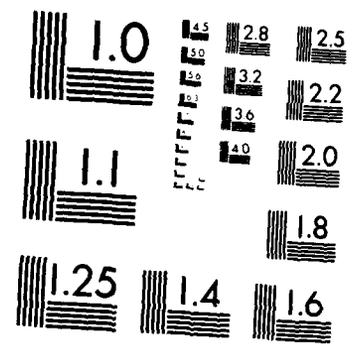
AD-A129 897 ROBUST STATISTICAL DATA ANALYSIS AND MODELING(U) TEXAS 1/1
A AND M UNIV COLLEGE STATION INST OF STATISTICS
E PARZEN MAY 83 ARO-16992.15-MA DAAG29-80-C-0070

UNCLASSIFIED F/G 12/1 NL





TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143

STATISTICS
Phone 409 - 845-3141

FINAL REPORT

TO ARMY RESEARCH OFFICE

ARO DAAG29-80-C-0070

"Robust Statistical Data Analysis and Modeling"

Professor Emanuel Parzen, Principal Investigator

May 1983

Texas A&M Research Foundation
Project No. 4226

4 February 1980 - 31 March 1983

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1 REPORT NUMBER	2 GOVT ACCESSION NO.	3 RECIPIENT'S CATALOG NUMBER
4 TITLE (and Subtitle) Final Report to Army Research Office		5 TYPE OF REPORT & PERIOD COVERED Technical 6 PERFORMING ORG. REPORT NUMBER
7 AUTHOR(s) Emanuel Parzen		8 CONTRACT OR GRANT NUMBER(s) ARO DAAG29-80-C-0070
9 PERFORMING ORGANIZATION NAME AND ADDRESS Texas A&M University Institute of Statistics College Station, TX 77843		10 PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11 CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12 REPORT DATE June 1983 13 NUMBER OF PAGES 15
14 MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15 SECURITY CLASS. (of this report) Unclassified 15a DECLASSIFICATION/DOWNGRADING SCHEDULE

16 DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17 DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18 SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

This is a final report on the research project entitled "Robust Statistical Data Analysis and Modeling" for the period 4 February 1980 - 31 March 1983.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

A. Goals and Achievements

The original proposal for this research project states its goals as follows: "The object of this research project is to continue a broad program of research whose aims are: (1) to develop the quantile and density-quantile function formulation of statistical data analysis and modeling problems; (2) to develop robust methods of data analysis and modeling; (3) to develop density estimation methods; and (4) to develop minimum distance methods and approximation theory methods. The aim of We propose to implement our theoretical research in algorithms and computer software which provide methods useable by researchers concerned with important scientific and social problems, and to discuss applications which illustrate the applicability of the methods."

The approach to statistical reasoning that our research program is attempting to develop has reached a synthesis that warrants its own name; we propose the name "FUN.STAT." The FUN.STAT domain of statistical data model identification and parameter estimation combines (1) density-quantile function signatures of distributions, (2) entropy and information measures, and (3) functional statistical inference.

The word "functional" is used with several interpretations: (a) functional = useful; (b) functional = functional analysis, as one applied techniques of numerical analysis, solutions of linear equations, and approximation theory; (c) functional = estimation of functions, and fitting curves and surfaces to a discrete grid of points. Functional inference is a branch of the field of "abstract inference" formulated by Grenander.

FUN.STAT is an approach to statistical graphics which argues that a graph should be a picture of a function (and the function should be a

signature of a probability model). FUN.STAT connotes the name of a library of computer programs for statistical data analysis whose output provides both graphs of functions and numerical diagnostics of the fit and complexity of the functions. We currently have available computer packages ONESAM, TWOSAM, and BISAM.

Some achievements of this research program are described in Section B which outlines the Quantile Data Analysis approach to one-sample, two-sample, and bivariate sample statistical data anlaysis problems. We believe that we have achieved important clarifications of the role of information and entropy measures in model identification and parameter estimation (information measures can be elegantly expressed in the quantile domain and appear to be more easily estimated in that domain).

Statistical concepts introduced or emphasized in our research include density-quantile function, quantile-density function, score function, tail exponents, mode percentile, sample quantile function, histogram-quantile function, quantile box plot, cumulative weighted spacings plot, sample entropy, score deviation, 19 quantile values for universal data summary, quantile bootstrap, joint density-quantile function, dependence density function, dependence entropy, regression-quantile function, Bayes theorem for quantile functions, autoregressive quantile densities, exponential dependence densities, minimum distance estimation by reproducing kernel Hilbert space norms, Renyi entropy of order α. These concepts seem to be increasingly accepted (and referred to) in the literature.

We believe that we have made excellent progress towards achieving the goals stated in our original proposal. A framework has been developed for integrating statistical data analysis and concepts of probability theory.

B. Summary of some of the most important results of Quantile and FUN.STAT Data Analysis

I. One Sample: Univariate

The probability law of a random variable X is usually described by its distribution function $F(x)=Pr[X\leq x]$, $-\infty<x<\infty$, and probability density function $f(x)=F'(x)$. The quantile approach uses

(1) $Q(u) = F^{-1}(u) = \inf \{x:F(x)\geq u\}$,

(2) $q(u) = Q'(u)$,

(3) $fQ(u) = f(Q(u)) = \{q(u)\}^{-1}$, and

(4) $J(u) = -(fQ)'(u)$.

A quick measure of location is the median Q(0.5). A quick index of scale is the interquartile range Q(0.75) - Q(0.25), formed from the quartiles Q(0.25) and Q(0.75).

Quick measures of distributional shape are provided by values (as u tends to 0 and 1) of the informative quantile function [recently introduced by Parzen].

$$IQ(u) = \frac{Q(u) - Q(0.5)}{2\{Q(0.75) - Q(0.25)\}} \quad , \quad 0\leq u\leq 1.$$

We cannot emphasize how powerful the IQ function appears to be in practice as a tool for the diagnosis of distributional shapes.

The IQ function is independent of location and scale parameters. It is approximately equivalent to normalizing a quantile function to have the properties Q(0.5) = 0, Q'(0.5) = 1. The IQ graph of the function provides us at a glance with a vague estimate of tail behavior as defined by tail exponents.

A fundamental description of the tail behavior of distributions is provided by the left tail exponent α_0 and the right tail exponent α_1 defined as follows:

$$fQ(u) = u^{\alpha_0} L_0(u) \text{ as } u \to 0$$

$$fQ(u) = (1-u)^{\alpha_1} L_1(u) \text{ as } u \to 1$$

where $L_0(u)$ and $L_1(u)$ are slowly varying functions.

A function $L(u)$ is slowly varying as $u \to 0$ if, for every $y > 0$,

$$\lim_{u \to 0} \frac{L(yu)}{L(u)} = 1.$$

Tail behavior is defined in terms of a tail exponent as follows:

$\alpha<1$: short tail

$\alpha=1$: medium tail

$\alpha>1$: long tail

Medium tail ($\alpha=1$) distributions are further classified by the value of

$$h_0 = \lim_{u \to 0} \frac{f(u)}{u} \quad , \quad h_1 = \lim_{u \to 1} \frac{f(u)}{1-u} \quad ;$$

the letter h is suggested by the notion of hazard function. We define

$h = 0$: medium-long tail

$0 < h < \infty$: medium-medium tail

$h = \infty$: medium-short tail .

Extensive calculations of informative quantile functions indicate that the value IQ_0 of $IQ(u)$ for u near 0 is a quick indicator of left tail behavior:

$-0.5 \leq IQ_0 < 0$: short left tail,

$-1.0 \leq IQ_0 < -0.5$: medium-short left tail,

$IQ_0 < -1.0$: medium-medium to long left tail.

Similarly the value IQ_1 of IQ(u) for u near 1 is a quick indicator of right tail behavior:

$0 < IQ_1 \leq 0.5$: short right tail,

$0.5 < IQ_1 < 1.0$: medium-short left tail,

$1.0 < IQ_1$: medium-medium to long right tail .

An important family of distributions is the Weibull with shape parameter β . Its quantile function Q(u) is of the form

$$Q(u) = \mu + \sigma Q_0(u)$$

where

$$Q_0(u) = \frac{1}{\beta} \{\log (1-u)^{-1}\}^{\beta} .$$

Its density-quantile

$$f_0 Q_0 (u) = (1-u) \{\log (1-u)^{-1}\}^{1-\beta} .$$

Its right tail exponent is $\alpha = 1.$, and its left tail exponent is $\alpha_0 = 1-\beta$. Insight into the interpretation of informative quantile functions is obtained by computing them for Weibull distributions.

Given data, we distinguish three types of estimators of population parameters, which we call: (1) fully non-parametric, (2) fully parametric, and (3) functional-parametric. Fully non-parametric estimators assume no model, and provide quick estimators. Fully

parametric estimators assume a model known up to a finite number of parameters which must be estimated. Functional-parametric estimators are based on methods of functional statistical inference.

A fully non-parametric estimator $\tilde{Q}(u)$ of $Q(j)$, given a sample of n distinct values $X_{1;n} < X_{2;n} < \ldots < X_{n;n}$, is defined by (for $j=1,\ldots,n$)

$$\tilde{Q}(u) = X_{j;n} \quad , \quad \frac{j-1}{n} < u \leq \frac{j}{n} \quad .$$

For a large sample, or for grouped values, we form a histogram before computing $\tilde{Q}(u)$ by linear interpolation at an equi-spaced grid of values kh, $k=1,2,\ldots,[1/h]$ where usually $h = 0.01$.

A quantile data analysis of the random sample

1. Forms sample distribution function $\tilde{F}_X(x)$, sample quantile function $\tilde{Q}_X(u)$, sample quantile density $\tilde{q}(u)$ at a grid of values of u in $0<u<1$.

2. Plots sample version of informative quantile function $IQ(u)$ whose values as u tends to 0 and 1 indicates the tail exponents of the probability law of X.

3. Determines standard distribution functions $F_0(x)$ to test

   $$H_0: F(x) = F_0\left(\frac{x-\mu}{\sigma}\right) \text{ or } Q(u) = \mu + \sigma Q_0(u)$$

   for location and scale parameters μ and σ to be estimated. A test of H_0 which does not require estimation of μ and σ can be based on [Parzen (1979)]

   $$\tilde{d}(u) = f_0 Q_0(u)\, \tilde{q}(u) \doteq \tilde{\sigma}_0 \quad ,$$

$$\tilde{\sigma}_o = \int_0^1 f_o Q_o(t)\, \tilde{q}(t)\, dt$$

which estimate respectively

$$d(u) = f_o Q_o(u)\, q(u) \div \sigma_o \quad ,$$

$$\sigma_o = \int_0^1 f_o Q_o(t)\, q(t)\, dt.$$

4. Form successive autoregressive estimators

$$\hat{d}_m(u) = \hat{K}_m \left| 1 + \hat{\alpha}_m(1)\, e^{2\pi i u} + \ldots + \hat{\alpha}_m(m)\, e^{2\pi i u m} \right|^{-2}$$

whose negentropy

$$\hat{H}_m = \int_0^1 - \log \hat{d}_m(u)\, du = - \log \hat{K}_m$$

is used to determine optimal orders $\hat{m}$. Note that $\hat{H}_m$ estimates the entropy difference

$$\Delta = \{\log \sigma_o - \int_0^1 \log f_o Q_o(u)\} - \{- \int_0^1 \log fQ(u)\, du\}$$

5. Estimate $fQ(u)$ by

$$\hat{fQ}_m(u) = f_o Q_o(u) \div \tilde{\sigma}_o\, \hat{d}_m(u)$$

where m is chosen equal to an optimal order $\hat{m}$.

II. Two Sample: Univariate

Let X and Y be continuous random variables with random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ respectively, and with respective distribution functions $F(x) = Pr[X \leq x]$, $G(x) = Pr[Y \leq x]$. The pooled sample $X_1, \ldots, X_m, Y_1 \ldots, Y_n$ can be regarded as a random sample from the distribution function

$$H(x) = \lambda F(x) + (1-\lambda) G(x), \qquad \lambda = \frac{m}{m+n} \quad .$$

To test the hypotheses of equality of distributions, H_o: $F(x) = G(x) = H(x)$, it is customary in non-parametric statistics to introduce

$$D_X(u) = F H^{-1}(u), \qquad D_Y(u) = G H^{-1}(u)$$

with densities [equivalent to likelihood ratios]

$$d_X(u) = \frac{f H^{-1}(u)}{f H^{-1}(u)} \quad , \qquad d_Y(u) = \frac{g H^{-1}(u)}{h H^{-1}(u)} \quad .$$

Note that $h H^{-1}(u) = \lambda f H^{-1}(u) + (1- \lambda) g H^{-1}(u)$; therefore

$$d_X(u) = \left\{ \lambda + (1-\lambda) \frac{g H^{-1}(u)}{f H^{-1}(u)} \right\}^{-1} \quad .$$

Parzen (1983) shows that all conventional two-sample nonparametric test procedures are functionals of the following raw estimator of $D_X(u)$:

$$\tilde{D}_X(u) = \{\tilde{H} \tilde{F}_X^{-1}\}^{-1}(u)$$

from which one can form "pseudo-correlations" $\tilde{\rho}(v)$ and linear rank statistics $\Delta(J)$ with score function $J(u)$,

$$\tilde{\rho}(v) = \int_0^1 e^{2\pi i u v} \, d \, \tilde{D}_X(u) \ , \quad \Delta(J) = \int_0^1 J(u) \, d\tilde{D}_X(u) \quad ,$$

and autoregressive estimators $\hat{d}_{X,m}(u)$ of $d_X(u)$.

When one observes several variables $X^{(1)}, X^{(2)}, \ldots, X^{(j)}, \cdots$ one estimates functionals of $D_j(u) = F_{X(j)} (H^{-1}(u))$ or $D_{jk}(u) = F_{X(j)}(F^{-1}_{X(k)}(u))$.

III. One Sample: Bivariate

Let (X_1, X_2) be jointly continuous random variables with distribution function $F_{X_1,X_2}(x_1,x_2) = Pr[X_1 \leq x, X_2 \leq x_2]$ and density $f_{X_1,X_2}(x_1,x_2)$. The joint density quantile function is defined by

$$fQ_{X_1,X_2}(u_1,u_2) = f_{X_1,X_2}(Q_{X_1}(u_1), Q_{X_2}(u_2)) \quad .$$

To estimate fQ we define

$$D_{X_1,X_2}(u_1,u_2) = F_{X_1,X_2}(Q_{X_1}(u_1), Q_{X_2}(u_2))$$

which is the distribution function of $U_1 = F_{X_1}(X_1)$, $U_2 = F_{X_2}(X_2)$; it has density

$$d_{X_1,X_2}(u_1,u_2) = \frac{\partial^2}{\partial u_1 \partial u_2} D(u_1,u_2)$$

satisfying

$$fQ_{X_1,X_2}(u_1,u_2) = fQ_{X_1}(u_1)\, fQ_{X_2}(u_2)\, d_{X_1,X_2}(u_1,u_2) \quad .$$

To estimate d_{X_1,X_2} from a random sample $(X_1^{(j)}, X_2^{(j)})$, $j=1,\ldots,n$, form

$$\tilde{D}_{X_1,X_2} = \tilde{F}_{X_1,X_2}(\tilde{Q}_{X_1}(u_1), \tilde{Q}_{X_2}(u_2))$$

and a raw estimator $\hat{d}_{X_1,X_2}(u_1,u_2)$. We smooth $\log \tilde{d}_{X_1,X_2}(u_1,u_2)$ by a smooth estimator $\log \hat{d}_{X_1,X_2}(u_1,u_2)$ minimizing a criterion similar to

$$\sum_{j=1}^{n} \left| \log d[U_1^{(j)}, U_2^{(j)}] - \log d_m[U_1^{(j)}, U_2^{(j)}] \right|^2$$

where $\log d_m(u_1,u_2)$ has the parametric representation (exponential model)

$$\log d_m(u_1,u_2) = \sum_{\nu_1,\nu_2} \theta_{\nu_1,\nu_2} \exp i\,(u_1\nu_1 + u_2\nu_2) - \psi(\theta_{\nu_1,\nu_2}) ;$$

where the summation is over $\nu_1,\nu_2 = 0, \pm 1,\ldots,\pm m$, and $\psi(\theta_{\nu_1,\nu_2})$ is an integrating factor to make $d_m(u_1,u_2)$ a probability density. The foregoing estimators have been implemented in T. J. Woodfield [1982]. The problem of choosing a best value of the order m is approached by evaluating the entropy of d_m.

References

Parzen, E. (1979) Nonparametric Statistical Data Modeling. Journal of the American Statistical Association, 74, 105-131.

Parzen, E. (1983) FUN.STAT Quantile Approach to Two Sample Statistical Data Analysis. The Canadian Journal of Statistics.

Woodfield, Terry Joe (1982) Statistical Modeling of Bivariate Data. Ph.D. Thesis Texas A&M Department of Statistics.

C. List of Publications and Technical Reports

The main publications by Professor Parzen on quantile data analysis and modeling are as follows:

Parzen, E. "Nonparametric Statistical Data Science: A Unified Approach Based on Density Estimation and Testing for White Noise" Technical Report, Statistical Science Division, SUNY at Buffalo, January 1977.

Parzen, E. "Nonparametric Statistical Data Modeling" Journal of the American Statistical Association, (with discussion), 74, 105-131, 1979.

Parzen, E. "A Density-Quantile Function Perspective on Robust Estimaton" Robustness in Statistics, ed. R. Launer and G. Wilkinson, New York: Academic Press, 237-258, 1979.

Parzen, E. "Density Quantile Estimation Approach to Statistical Data Modeling", Smoothing Techniques for Curve Estimation, ed. T. Gasser and M. Rosenblatt, Heidlberg: Springer, Lecture in Mathematics, 757, 155-180, 1979.

Parzen, E. "Comments on Good and Gaskins 'Density Estimation and Bump Hurting ...'", Journal of the American Statistical Association, 75, 56-59, 1980.

Parzen, E. "Quantile Functions, Convergence in Quantile, and Extreme Value Distribution Theory," Technical Report B-3, Texas A&M Institute of Statistics, November 1980.

Parzen, E. Comments on "Nonparametric standard errors and confidence intervals" by Bradley Efron, Canadian J. Statistics, 9, 164-165, 1981.

Parzen, E. "Data Modeling Using Quantile and Density-Quantile Functions", Proceedings of 1980 Lisbon Academy of Sciences Symposium on Recent Advances in Statistics. Academic Press: New York, 1982.

Parzen, E. "Quantiles, Parametric-Select Density Estimation, and Bi-Information Parameter Estimators," Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface Ed. K. W. Heiner, R. S. Sacher, J. W. Wilkinson. Springer Verlag; New York, 1983, 241-245.

Parzen, E. "Entropy Interpretation of Goodness of Fit Tests" Proceedings of the 28th Army Conference on the Design of Experiments, October 1982, Monterey, California.

Parzen, E. FUN.STAT Quantile Approach to Two Sample Statistical Data Analysis. The Canadian Journal of Statistics, December 1983.

Technical Reports

Contract DAAG29-80-C-0070

No.; Date	Author(s)	Title
B-1 March, 1980	Emanuel Parzen and Scott Anderson	ONESAM, A Computer Program for Non-parametric Data Analysis and Density Quantile Estimation
B-2 April, 1980	Emanuel Parzen	Data Modeling Using Quantile and Density-quantile functions
B-3 November 1980	Emanuel Parzen	Quantile Functions, Convergence in quantile, and extreme value distribution theory
B-4 November 1980	James Michael White	A quantile function approach to the K-sample quantile regression problem
B-5 April 1981	Thomas J. Prihoda	A Generalized Approach to the Two Sample Problem: The Quantile Approach
B-6 June 1982	Emanuel Parzen	Quantiles, Parametric-Select Density Estimation, and Bi-Information Parameter Estimators
B-7 August 1982	Terry Joe Woodfield	Statistical Modeling of Bivariate Data
B-8 January 1983	Emanuel Parzen	Entropy Interpretation of Goodness of Fit Tests

D. Ph.D. Theses

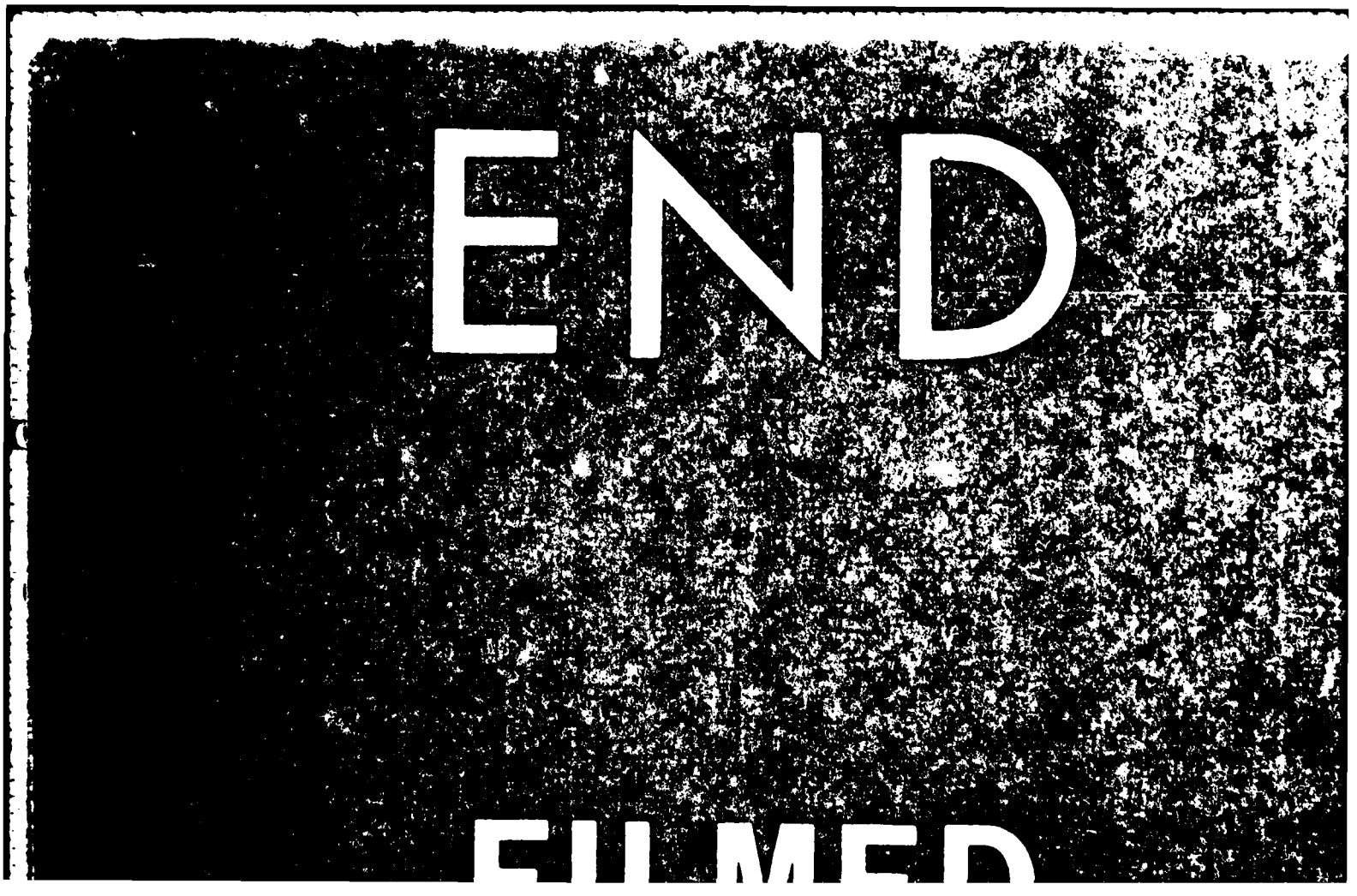
Four Ph.D. theses under Professor Parzen's direction have been completed with support from the Army Research Office in the years of 1979-1982. The theses of R. L. Eubank, J. M. White, T. J. Prihoda, and T. J. Woodfield focused respectively on the quantile and density-quantile approach to estimation of location and scale parameters; comparison of k samples; estimation of location and scale differences of two samples; and estimation of bivariate joint density-quantile functions. The work of S. Anderson was unfortunately terminated in 1982 by his accidental death.

Eubank, R. L. "A Density-Quantile Function Approach to Choosing Order Statistics for the Estimation of Location and Scale Parameters" Technical Report A-10, Texas A&M, Institute of Statistics, July 1979.

Prihoda, Thomas J. "A Generalized Approach to the Two Sample Problem: The Quantile Approach", Technical Report B-5, Texas A&M, Institute of Statistics, April 1981.

White, James Michael "A Quantile Function Approach to the K-Sample Quantile Regression Problem", Technical Report B-4, Texas A&M, Institute of Statistics, November 1980.

Woodfield, Terry J. "Bivariate Modeling of Bivariate Data", Technical Report B-7, Texas A&M, Institute of Statistics, August 1982.


END
FILMED